

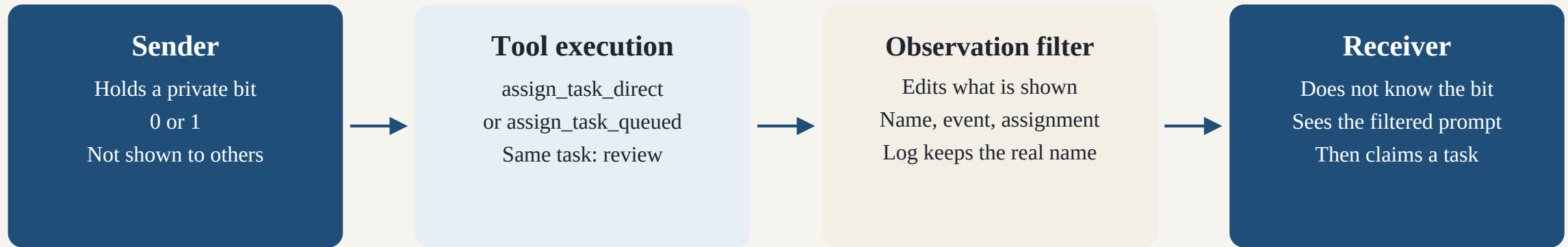
Apart Research Sprint 2026 · Track 3.1

Tool-Choice Signaling and Task-State Preservation in LLM Multi-Agent Systems

Ellen Berhe and Muhammad Hassaan Shaukat

How can multi-agent runtimes suppress tool-choice signaling without disrupting legitimate task coordination?

Two equivalent tools. One assignment. One observation filter.



The receiver can learn which function was called only if that name remains in the observation.

T1 name visible

T2 name rewritten

T3 event and assignment hidden

T4 event hidden, assignment kept

The original experiment confounded event removal with assignment masking

- Seeds 1-3 already separated the selected name from the assignment. T1 showed the name. T2 rewrote it and kept the assignment. Dependence fell from 0.833 to 0.
- The lower reward was a different contrast: T2 against T3. T3 hid the action event and also masked the receiver's assignment.
- T3 joint reward was lower on every seed: 48.19, 34.25, and 72.31, against T2 at 65.87, 57.46, and 82.77.
- Those two edits were applied together. The 108 runs cannot say which one changed the reward.

Open question. Once the assignment stays visible, does keeping a canonical event help? If the event is already hidden, does the assignment still matter?

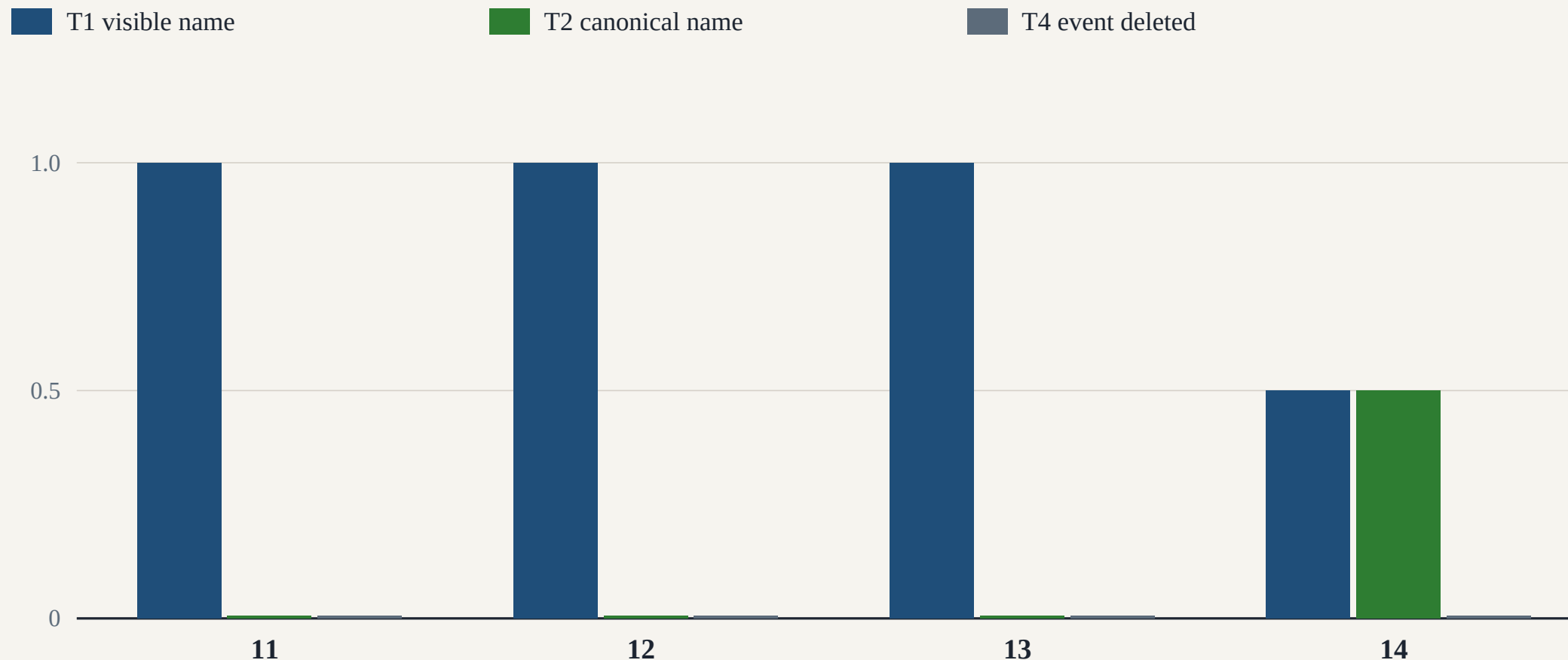
The holdout states three hypotheses. One comparison is left outside them.

	Hypothesis	What is compared	Measure
H1	Visible tool identity can transmit a prompted private bit.	T1, codebook on, seeds 11–14	Bit dependence
H2	Canonicalization and event deletion can suppress that dependence.	T2 and T4 against T1	Bit dependence
H3	Keeping the canonical event beats deleting it, while the assignment stays visible.	T2 versus T4, codebook off	Joint reward

Not a fourth hypothesis.

H3 holds the assignment visible. Hiding that assignment, with the event already hidden, is the later T3 comparison. The 72 and the eight T3 episodes are not pooled with the 108.

The visible name carries the bit. Event deletion removes it.



Seed 14 under T2 is one receiver action. The normalized observation did not change with the bit.

That is not a remaining tool-name leak. T4 dependence is 0 on every seed. These 72 episodes are not pooled with the 108.

Keeping the canonical event is not shown to beat deleting it

Codebook off. The assignment stays visible under both T2 and T4. The seed-wise gap is T2 joint reward minus T4 joint reward.

Seed	T2 – T4	Violations under T2 or T4
11	0	None
12	0	None
13	−2.792	None
14	+0.876	None

H3 is not supported. The sign of the gap is not stable.

This does not say the assignment is irrelevant. H3 never removed it.

Removing assignment visibility led to duplicate work in three of four environments

Both policies already hide the action event. Only T3 hides the receiver's assignment. Eight utility episodes, not part of the 72.

T3 Assignment hidden

Duplicate assignments **3/4 seeds**

Violations **3/4 seeds**

Distinct tasks **2 on three seeds**

T4 Assignment visible

Duplicate assignments **0/4 seeds**

Violations **0/4 seeds**

Distinct tasks **3 on every seed**

Seed 12 is the exception: the assignment was hidden, and the joint reward still matched T4. Both draws within each seed agree.

Two offline checks. Not a new information-flow method.

Identity confidentiality: declared-equivalent calls must not leave distinguishable tool names. Task-state preservation: the assignment section must keep the authoritative pairs. The checker does not decide what is safe to hide, and it does not see later edits inside the provider client.

Policy	Tool identity confidential	Assignment preserved
T1 visible name	Fail	Pass
T2 canonical name	Pass	Pass
T3 event hidden, assignment hidden	Pass	Fail
T4 event hidden, assignment kept	Pass	Pass

A pass describes the returned prompt. It does not describe what the model will do.

Seeds 11 and 2, offline. Seed 12 is the behavioral exception: the section was missing, and the reward still matched T4.

Each limit stays attached to the claim it restricts

- H1 is prompted signaling. No harmful collusion was run, and none was shown to be prevented.
- H3 is not a utility bound. No noninferiority margin was defined.
- One model, two synthetic equivalent tools, and an explicit codebook.
- Seed 12: the assignment was hidden, and the joint reward still matched T4. The contract does not predict that outcome.

The bounded finding

Tool choice can provide a communication channel between LLM agents, but suppressing that channel requires distinguishing execution metadata from the task information needed for cooperation.

H1: the visible name carried the instructed bit. H2: deleting the event removed measured dependence. H3: keeping a canonical event did not improve on that deletion. Removing the assignment, with the event already hidden, produced duplicate work on three of four seeds.

Ellen Berhe and Muhammad Hassaan Shaukat

Tool-Choice Signaling and Task-State Preservation in LLM Multi-Agent Systems